AL-CR-1992-0001

ARMSTRONG LABORATORY

ON THE EFFECT OF RANGE RESTRICTION ON CORRELATION COEFFICIENT ESTIMATION

Douglas E. Jackson

Eastern New Mexico University
Portales, NM 88130

University Energy Systems (UES)
Suite 600, 8961 Tesoro Drive
San Antonio, TX 78217

Malcolm James Ree

HUMAN RESOURCES DIRECTORATE
MANPOWER AND PERSONNEL RESEARCH DIVISION
Brooks Air Force Base, TX 78235-5000

April 1992

Interim Contractor Report for Period 1 January 1991 – 31 December 1991

Approved for public release; distribution is unlimited.

AIR FORCE SYSTEMS COMMAND
BROOKS AIR FORCE BASE, TEXAS 78235-5000

NOTICES

This contractor report is published as received and has not been edited by the technical editing staff of the Armstrong Laboratory.

Publication of this report does not constitute approval or disapproval of the ideas or findings. It is published in the interest of scientific and technical information (STINFO) exchange.

When Government drawings, specifications, or other data are used for any purpose other than in connection with a definitely Government-related procurement, the United States Government incurs no responsibility or any obligation whatsoever. The fact that the Government may have formulated or in any way supplied the said drawings, specifications, or other data, is not to be regarded by implication, or otherwise in any manner construed, as licensing the holder, or any other person or corporation; or as conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

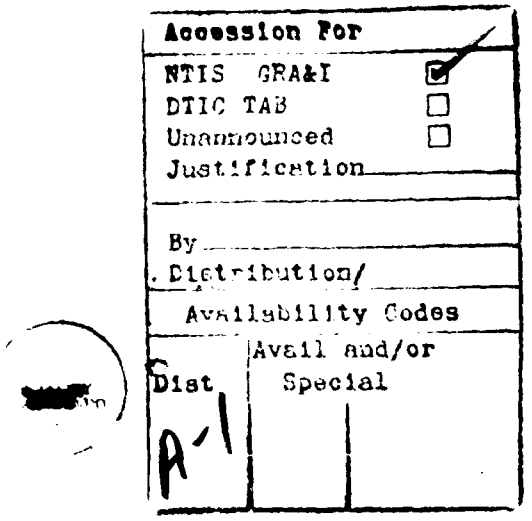
The Office of Public Affairs has reviewed this report, and it is releasable to the National Technical Information Service, where it will be available to the general public, including foreign nationals.

This report has been reviewed and is approved for publication.

MALCOLM JAMES REE
Senior Scientist

WILLIAM E. ALLEY, Technical Director
Manpower and Personnel Research Division

ROGER W. ALFORD, Lt Colonel, USAF
Chief, Manpower and Personnel Research Division



April 1992 | Final 1 January 1991 – 31 December 1991

F49620-88-C-0053
PE - 62205F
PR - 7719
TA - 18
WU - 67

On the Effect of Range Restriction on Correlation Coefficient Estimation

Douglas E. Jackson
Malcolm James Ree

Universal Energy Systems (UES)
Suite 600, 8961 Tesoro Drive
San Antonio, TX 78217

Armstrong Laboratory
Human Resources Directorate
Manpower and Personnel Research Division
Brooks Air Force Base, TX 78235-5000

AL-CR-1992-0001

Armstrong Laboratory Technical Monitor: Malcolm James Ree, (512) 536-3256.

Approved for public release; distribution is unlimited.

Suppose it is desired to estimate the correlation coefficient between random variables X and Y in some population P and the only data available are from some population Q where Q is a proper subset of P. X and Y are defined on P, while X^* and Y^* will denote X and Y restricted to Q. A simulation program was written to study the effect of this restricted sampling on the estimation of correlation coefficients. The Air Force is studying the implementation of new selection devices that optimize the selection and classification of individuals. Whenever a new measurement instrument is suggested, it must be evaluated by estimating its correlation with performance criteria and with tests and selection devices that are currently part of the selection process. The difficulty is that the new test can only be administered to Air Force personnel. That is, people who have already been selected. Air Force personnel constitute the population Q and the applicants constitute the population P. It is necessary to use a sample from Q to estimate correlations between tests that are to be used in P. This is called the range restriction problem. The purpose of this paper is to present the results of a study which addresses a number of issues related to the range restriction problem. The performance of the F-statistic, confidence intervals, and "hidden variables" are considered.

Personnel tests | Training performance | 46
Pilot candidate selection | Undergraduate pilot training

Unclassified | Unclassified | Unclassified | UL

CONTENTS

APPENDICES

Appendix

LIST OF TABLES

Table No.

FIGURES

Figure

PREFACE

This report summarizes an investigation of the correction for range restriction. It was conducted under Air Force contract number F49620-88-C-0053 and is the final report from that contract. It provides an investigation of a critical tool which will help the Air Force in its search for and understanding of new predictors and predictor-criterion relationships. The authors wish to thank Dr. L. D. Valentine, Jr., Dr. William Alley and Colonel Daniel Leighton of the Manpower and Personnel Research Division of the Human Resources Directorate of the Armstrong Laboratory.

SUMMARY

The Air Force is considering new tests for selection and classification of both enlisted and commissioned entrants. The new tests are typically evaluated in samples which have been preselected on the basis of some related measures which cause an attenuation of the observed validity correlations. This paper reports on the performance of the F test statistics and confidence interval as well as the problem of hidden variables in making decisions about these correlations.

ON THE EFFECT OF RANGE RESTRICTION ON COEFFICIENT ESTIMATION

I. INTRODUCTION

When certain linearity and homoscedasticity conditions are satisfied, there is a theorem that shows how ρ_{XY} (the correlation in P) may be computed from $\rho_{X^*Y^*}$ the correlation in Q). The result was first demonstrated by Pearson (1903) and then strengthened by Lawley (1943). $r_{X^*Y^*}$, which is calculated using a sample from Q, is an estimate of $\rho_{X^*Y^*}$, and Pearson's formula may be used to compute an estimate of ρ_{XY} by using $r_{X^*Y^*}$ in the place of $\rho_{X^*Y^*}$. This estimate is sometimes called the corrected correlation coefficient or Pearson's correction statistic or simply the correction statistic. A simulation program (Jackson & Ree, 1990) was written to evaluate the correction statistic and it was found to work very well when the joint distribution of all tests is multinormal. The current study investigated a number of questions related to the correction. This section contains a list of these questions and a statement of Lawley's theorem.

When a new test is a candidate for inclusion in an enlistment qualification battery or other system, a standard F test is performed to decide if the new test adds to the prediction of the system. An obvious question is whether restricted sampling might bias this F test and whether correction might bias the F test. Section II is devoted to this question.

There are certain variables that influence personnel selection that are not part of the test battery, and hence, are not included in the calculation of the correction statistic. The reason might be that the variable is not known or that it is difficult to measure. This is referred to as the "unknown variable" problem. In Section III, the mathematical reasons that unknown variables degrade the accuracy of the correction statistic is investigated. The magnitude of this degradation is studied by simulation, and one solution proposed.

The Fisher z-transformation (Z-transform) of the corrected correlation coefficient between X and Y appears to have normal distribution when X and Y come from a bivariate normal distribution. Whereas it is provable that the z-transform of the ordinary sample correlation coefficient (r_{XY}) has a normal distribution, only empirical evidence of the analogous result for the corrected statistic can be given. Evidence is also presented that the mean of the z-transform of the corrected statistic is very close to the z-transform of ρ_{XY} in the

multinormal case. In other words, the inverse Z-transform applied to the mean of the Z-transforms of a random sample of correction statistic observations is ρ_{XY}. These observations would lead to a method of calculating confidence intervals for ρ_{XY} if only the variance of the z-transform of the corrected statistic were known. In order to obtain estimates of this variance, a simulation program has been written. These matters are addressed in Sections IV and V.

Up to now, all comparisons of the uncorrected versus the corrected statistic has been made for multinormal distributions. The data to which these estimates are applied may not be multinormal or indeed even linear and homoscedastic. In Section VI the uncorrected and the corrected statistic are compared on real data for Air Force enlisted members. These data were 3,930 test records, where each record had 11 test variables. The 11 variables were the 10 tests of the Armed Services Vocational Aptitude Battery (ASVAB) (Ree, Mullins, Mathews, 1982) and one criterion score.

It is appropriate to include in this section a statement of the correction formula and the set of minimal assumptions necessary for its application. The following theorem is due to Lawley (1943). Variables that are part of the selection criteria are called explicit selection variables and all others are called incidental selection variables. P is the applicant group and Q is the selected or restricted group.

Let X be the p-element vector of explicit selection variables, and Y the n - p element vector of incidental selection variables in the applicant group. Then X^* and Y^* represent the explicit and incidental selection variables in the selected group. Let

$$V = \begin{bmatrix} V_{p,p} & V_{p,n-p} \\ V_{n-p,p} & V_{n-p,n-p} \end{bmatrix}$$

represent the variance-covariance matrix for X*, Y*. The first p rows and columns refer to the components of X^*. So $V_{p,p}$ is the variance-covariance matrix of X^*, $V_{n-p,n-p}$ is the variance-covariance matrix for Y^*, $V_{p,n-p}$ gives the covariances between X^* and Y^*, and $V_{n-p,p}$ is the transpose of $V_{p,n-p}$. In this discussion V refers to selected data and W refers to applicant (or unselected) data. V will be the estimates of the variance-covariance of all tests and it is based on selected data. The restricted population consist of those who were accepted into the

organization and there are data on all tests for these people. Let

$$W = \begin{bmatrix} W_{p,p} & W_{p,n-p} \\ W_{n-p,p} & W_{n-p,n-p} \end{bmatrix}$$

be the matrix of variance-covariances for the applicant data. $W_{p,p}$ will be estimated from the data since there are data for the explicit selection variables on all applicants. The $W_{p,n-p}$, $W_{n-p,p}$, and $W_{n-p,n-p}$ are the matrices to be calculated and will be given to us by the theorem. $W_{n-p,p}$ is, of course, the transpose of $W_{p,n-p}$ so only the expression for $W_{p,n-p}$ will be given when the theorem is stated. The following statement of the theorem is taken from Birnbaum, Paulson, and Andrews (1950).

Assumption 1: (Linearity) For each j, the true regression of Y_j on X is linear.

Assumption 2: (Homoscedasticity) The conditional variance-covariance matrix of Y given X does not depend on X.

Theorem: Under assumptions 1 and 2

$$W_{p,n-p} = W_{p,p} V_{p,p}^{-1} V_{p,n-p} \quad \textit{and}$$

$$W_{n-p,n-p} = V_{n-p,n-p}, - V_{n-p,p} \left(V_{p,p}^{-1} - V_{p,p}^{-1} W_{p,p} V_{p,p}^{-1} \right) V_{p,n-p}$$

If sample correlation coefficients are used instead of population parameters in the matrices $W_{p,p}$, $V_{p,p}$, $V_{p,n-p}$, $V_{n-p,p}$ and $V_{n-p,n-p}$ then the entries of the matrices $W_{p,n-p}$, and $W_{n-p,n-p}$ are estimates of the correlations in the applicant population. These estimates are the corrected statistics of interest and a simulation program CORR was written to study their sampling distribution. Appendix A contains a general description of this program and Appendix B is a reference manual for its use.

II. THE F TEST

Since the general considerations are not significantly different from the one versus two independent variables model, only this special case is treated. It is assumed here that Y is the criterion variable, X_1 is the only explicit selection variable, and X_2 is a candidate to become one if it increases prediction of which individuals will have high Y scores. E is used to indicate error. In the full model

and in the reduced model $$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + E_f$$

$$Y = \beta_0' + \beta_1' X_1 + E_r.$$

It is assumed in the model that

$$E(E_f|\chi_1,\chi_2) = E(E_r|\chi_1) = 0,$$

$$Var(E_f|\chi_1,\chi_2) = \sigma^2_{E_f},$$

and

$$Var(E_r|\chi_1) = \sigma^2_{E_r}.$$

In other words, the mean of E for given X values is zero and the variance of E for given X values does not depend on those X values. It is also assumed that the distribution of E for any given set of X values is normal and is independent of the distribution of E for any other set of X values. A discussion of the F test may be found in any standard text that covers multiple regression, for example Dunn and Clark (1974).

The null hypothesis for this test is

$$H_0 : \beta_2 = 0.$$

The test statistic is

$$F = \frac{(SSE_{r*} - SSE_{f*})/((n-2)-(n-3))}{SSE_{f*}/(n-3)}$$

where $SSE_{r*}(SSE_{f*})$ is the sum of the squares due to error for the reduced (full) model and n is the sample size. The sum of squares due to error is the sum of the squares of the vertical distances between the individual data points and the corresponding points on the best least squares regression line or plane. The * characters indicate that the samples used to calculate these statistics are taken from the restricted population. This is necessary, of course, due to

the fact that for the X_2 and Y variables, only restricted data are available. Under the assumptions on the full model, plus the null hypothesis, the sampling distribution of F is an F-distribution with 1 numerator degrees of freedom and n - 3 denominator degrees of freedom. Notice that no assumptions are necessary concerning the distribution of X_1 or X_2. It is only required that E for fixed values of X values is normal, and the distributions of E for fixed values of X are independent. See Chatterjee and Price (1977) for a statement of this result. These conditions are satisfied in the restricted population if they are satisfied in the applicant population. However, it is necessary to show in addition that applying the correction statistic to the F calculation is and identity operation. It has no effect at all. But first we need to explain what is meant by applying the correction formula to the F calculation.

The F statistic is usually given in terms of sums of squares. However, it may also be written in terms of multiple correlation coefficients, and thus it is appropriate to consider the application of the correction formula to these correlations. It is observed that the correction formula works well on the multiple correlation coefficients but has no effect on the value of F.

Now define

$$\hat{Y}_{12} = \beta_0 + \beta_1 X_1 + \beta_2 X_2,$$

$$\hat{Y}_1 = \beta_0' + \beta_1' X_1,$$

$$R_f^2 = \rho_{y\hat{y}_{12}} \text{ and } R_r^2 = \rho_{y\hat{y}_1}.$$

R_f^2 and are called multiple correlation coefficients. Under the assumptions of the full model and the null hypothesis, the conditions of Lawley's theorem are met and the correction formula may be applied to estimates of R_{r*}^2 and R_{f*}^2 to obtain estimates of R_f^2 and R_r^2.

Let $S_{f*}^2\left[S_{r*}^2\right]$ be the standard sample statistic for estimating $R_{f*}^2\left[R_{r*}^2\right]$. That is, let $S_{f*}^2\left[S_{R*}^2\right]$ be the sample correlation coefficient between Y^* and $\hat{Y}_{12}^*[Y^* \text{ and } \hat{Y}_1^*]$. The reader should realize that in the previous sentence $\hat{Y}_{12}^*\left[\hat{Y}_1^*\right]$ is actually defined using the sample estimates of β_0, β_1, and β_2 $[\beta_0', \text{ and } \beta_1']$. Again * indicates the restricted

population.

It is easy to show that

$$(n-1)(1-S_{f*}^2)S_{Y*}^2 = SSE_{f*}$$

and

$$(n-1)(1-S_{r*}^2)S_{Y*}^2 = SSE_{r*}.$$

Then a simple algebraic manipulation gives

$$F = \frac{S_{f*}^2 - S_{r*}^2}{1-S_{f*}^2}(n-3).$$

The claim is that this value of F does not change if the corrected values of S_{f*}^2 *and* S_{r*}^2. are used in this formula instead of applying it as written. Now the correction formula applies to population parameters instead of sample statistics, and hence, it must be shown that

$$\frac{R_{f*}^2 - R_{*}^2}{1-R_{f*}^2} = \frac{R_F^2 - R_r^2}{1-R_f^2}.(1)$$

Corrections, of course, are calculated using sample estimates, while the correction formula express a relationship holding for population parameters. But any estimates based on this formula will satisfy the same relationships as those holding for population parameters. It follows immediately that in both the full and reduced models, the variance of the error terms are the same in the applicant population as in the restricted population. It is well known that $(1 - R^2)\ \sigma_Y^2 = \sigma_E^2$ in both the applicant and the restricted population. In some texts, for example Dunn and Clark (1974), the equation in the last sentence is taken as the definition of the multiple correlation coefficient. Our definition is equivalent. Hence

$$(1-R_f^2)\sigma_y^2 = \sigma_{E_f^2} = \sigma_{E_{f*}}^2 = (1-R_{f*}^2)\sigma_{Y*}^2$$

and

$$(1-R_r^2)\sigma_y^2=\sigma_{E_r^2} = \sigma_{E_{r*}}^2 = (1-R_{r*}^2)\sigma_{Y*}^2.$$

But equation 1 follows from these two equations by the same manipulations used above to derive a formula for F in terms of the sample multiple correlation coefficients.

Program CORR was modified to calculate the F statistic in the two versus one variable case. In the following example the program calculated 100 F values, each based on a sample of size 63. Each variable has a mean of zero and a standard deviation of one. The three variables are X_1,X_2, and Y with $\rho_{X_1,X_2} = .707$, $\rho_{X_1,Y} = .707$, and $\rho_{X_2Y} = .5$. The formula for β_2 is

$$\beta_2 = \frac{\rho_{X_2,Y} - \rho_{Y,X_1}\rho_{X_1,X_2}}{1 - \rho_{X_1,X_2}^2}\left(\frac{\sigma_Y}{\sigma_{X_2}}\right)$$

which is zero in this case and so the null hypothesis is satisfied. Four runs were made and for each, the mean of the 100 F values is given as well as the number of values exceeding 4.0, which is the .05 critical value for an F distribution with 1 numerator and 60 denominator degrees of freedom. The expected value of this F distribution is 1.03. The means of the 100 F values in the four runs were .904, 1.236, .991, and 1.122. The number of F values 4.0 or larger in the four runs were 5, 7, 5, and 5. Next, four runs were made with the same parameters except that the population was restricted to those observations having $X_1 \geq 0.67$. Since X_1 is a standard normal random variable, this yields a selection ration of 0.25. For these four runs, the means of the 100 F values were 1.177, .981, .997, and .964. The number of F values 4.0 or greater were 6, 6, 5, and 4. These empirical observations tend to confirm the conclusion that the F statistic is not affected by range restriction.

III. HIDDEN VARIABLES

Consider the effect of an explicit selection variable that was not included in the calculation of the corrected correlation coefficient using Lawley's theorem. The reason for exclusion of this variable might be that it was unknown to the individual doing the correction.

or perhaps difficult to measure. Assuming that the hypotheses of Lawley's theorem would be satisfied if all explicit selection variables were included, then, they most likely will not be satisfied if one were omitted. This effect is easily observable when there are only three variables and so all examples include just three variables. First, it is shown mathematically why the correction formula should be expected to fail and then a number of simulations are presented to give an idea of the magnitude of the inaccuracies caused by hidden variables. The reference manual for CORR in Appendix B includes a discussion of how to use this feature of the program.

Consider the model

$$(1) \qquad Y = \beta_0 + \beta_1 X_1 + E_1,$$

where X_1 and E_1 are quasi independent. Quasi independent means that the mean of E_1 for any given value of X_1 is zero, and the variance of E_1 for any given value of X_1 does not depend on that given value. Under these circumstances, it follows immediately from the definition of covariance that

$$cov(X_1, E_1) = 0.$$

Now if the above assumptions hold, and X_1 is the only explicit selection variable, then the correction formula can be used to estimate the correlation between X_1 and Y in the applicant population. Suppose, however, that there is another explicit selection variable, X_2, and that

$$(2) \qquad Y = \beta_0' + \beta_1' X_1 + \beta_1' X_2 + E_{12},$$

where E_{12} is quasi independent of X_1 and X_2. This last statement just means that the mean of E_{12} for given values of X_1 and X_2 is zero and that the variance of E_{12} for given values of X_1 and X_2 does not depend on those values. Again, it follows that

$$cov(X_1, E_{12}) = cov(X_2, E_{12} = 0.$$

The problem with applying Lawley's theorem to Model 1, when X_2 is an explicit selection variable, is that E_1 may not be quasi independent of X_2, which is required by Lawley's theorem since X_1 and X_2 are both explicit selection variables. We are assuming that E_{12} is

quasi independent of X_1 and X_2 but that does not imply that E_1 is quasi independent of these two random variables. Indeed, it can be shown that if Model 1 holds, then

$$cov(X_2, E_1) = cov(X_2, Y) - \frac{cov(X_1, Y)cov(X_1, X_2)}{\sigma_1^2}$$

and if this quantity does not happen to be zero, then we know that X_2 and E_1 are not quasi independent for that would imply that $cov(X_2, E_1) = 0$. So if Model 1 and 2 both hold, as they do for the multinormal distribution, and X_2 is not included in the correction calculation, then the corrected values will most likely be wrong. It is exactly this scenario that was assumed in the following simulations.

The examples presented here were chosen to demonstrate that the effect of variables missing from the correction calculation can be significant. They are not presented as typical examples. The three variables are Y, X_1, and X_2, where each is a standard normal, X_1 is the known explicit selection variable, and X_2 is the hidden explicit selection variable. The correction procedure for one explicit selection variable is used ($X_1 \geq 0$) but in reality the selection criteria are

$$X_1 \geq 0$$

and

$$X_2 \geq 0.$$

In the first example, the unrestricted population parameters are $\rho_{X_1,Y} = .5, \rho_{X_1,X_2} = .8$, *and* $\rho_{X_1,Y} = 0$. $\rho_{X_2,Y}$ is the parameter being estimated. A sample size of $n = 200$ was used. Thus, one observation of the corrected statistic involves generating multinormal observations until 200 have satisfied the selection criteria and then using this data in the correction formula. One run of CORR calculates the corrected statistic 100 times and displays the distribution of these values and several sample statistics including the mean and standard deviation of the 100 values. Two runs produced a mean corrected sample statistic of .677 for the first and .682 for the second. The standard deviation for both runs was .05. The selection ratios were .398 and .402 for the two runs. Recall that the true value of $\rho_{X_1,Y}$ is .5 and hence, we clearly have an overestimate of somewhere between 35 and 37 percent.

Now the hidden variable is removed and the two selection criteria are replaced by the one $X_1 \geq .25$. The value .25 was chosen to produce a selection ratio of .4 so as to be comparable with the previous runs. Two runs produce mean corrected sample statistics of .504 and .509. The standard deviations were .08 for the first run and .09 for the second. The selection ration was .401 for both runs. Thus, it is clear that the significant overestimates of the previous two runs were caused by the hidden variable.

The following is an example showing an underestimate caused by a hidden variable. This time, it is assumed that $\rho_{X_1,Y} = .3$, $\rho_{X_1,X_2} = .4$, and $\rho_{X_2,Y} = .8$. X_1 is the known

and X_2 is the unknown explicit selection variable and the selection criteria are again

$$X_1 \geq 0$$

and

$$X_2 \geq 0.$$

The sample size is 200 and the parameter being estimated is $\rho_{X_1,Y} = .3$. Two runs were made and the mean corrected statistics were .160 and .174. The standard deviation of the corrected sample statistic was .11 in both cases. This means that the standard deviation of the estimate of the mean corrected statistic, based on 100 repetitions, is approximately .011. Hence, there is a significant underestimate in the range of 42 to 47 percent. The selection ratio was .315 for both runs.

Again, the hidden variable is removed by replacing the two selection criteria by

$$X_1 \geq .48.$$

These two runs gave a mean corrected statistic of .297 and .284. The standard deviation of our estimate of the mean corrected statistic was approximately .012 for both runs. The selection ratios were .314 and .316. The underestimate of the previous two runs was clearly caused by the hidden variable.

Significant hidden variable effects are clearly possible. Whether or not significant inaccuracies exist in real applications is not known. The following parameter values were taken from the Air Force data based on test scores. Two cases were considered, one with, and one without a hidden variable just as in the previous examples. This example is less contrived than the last two, which were deliberately chosen to produce dramatic results. The

details of the present case are exactly as in all previous cases except that $\rho_{X_1,Y} = .71, \rho_{X_2,Y} = .7,$ and $\rho_{X_1,X_2} = .83.$ The mean corrected statistic for the 100 replications was .677 for one run and .678 for the other. The standard deviation of our estimate of the mean corrected statistic was about .006 for both runs. The selection ratio was .4 both times. This slight downward bias in the estimate of $\rho_{X_1,Y} = .71$ is significant but not severe.

Now the hidden variable is removed just as before by replacing the two selection criteria by one condition on X_1 selected to produce a selection ratio of .4. The two estimates for these runs were .697 and .706 and the standard deviation of our estimate was about .005 in both runs. The hidden variable caused a slight downward bias.

Solutions to the hidden variable problem have been proposed, for example, by Gross and McGanney (1987). They assume a slightly different model. In their model, the selection condition is a single inequality stating that a linear combination of the observable explicit selection variables plus an error term is nonnegative. The error term plays the role of the hidden variable or variables. In the current model, the criteria consists of several equations which all must be satisfied and one of these equations involves an unobservable or hidden variable. The Gross McGanney model assumes that

$$Y = \beta_0 + \beta_1 X + E_y$$

and

$$Y_s = \alpha_0 + \alpha' X_s + E_s,$$

where $\alpha' X_s$ denotes a linear combination of the observable explicit selection variables and $Y_s \geq 0$ is the selection condition. Also, X and E_y, as well as X_s and E_s, are required to be quasi independent as previously defined. Finally, it is assumed that E_y given X and E_s given X_s are both normally distributed. The authors display a number of simulations based on a maximum likelihood estimation of the parameters in the model. They do not have an analytic equation for the maximum likelihood estimators but rather use the Newton-Raphson numerical technique to find the maximum of the likelihood function. This method works well when the conditions of the model are met and ρ_{Y,Y_s} is not close to zero. When ρ_{Y,Y_s} is close to

zero, the Pearson correction formula works well. They decide which statistic to use by doing a hypothesis test for $\rho_{Y,Y_s} = 0$. This test, of course, depends on the assumption that the error terms are normal.

There are a number of objections to using this maximum likelihood estimator. Many assumptions are necessary for its use. The selection conditions must all fit into one equation and the error terms must be normal. Both of these assumptions are false for most selection testing data. In addition, it is not known for certain that significant inaccuracies are caused by hidden variables in practice. The Pearson correction, on the other hand, requires no assumption of normality or constraints on the selection criteria.

The assumptions required to use Pearson correction are not known to hold perfectly in the absence of hidden variables.. In the simulations presented earlier in this section, there is that implicit assumption that they do hold in the absence of hidden variables. If this assumption were true, then the most prudent course would be to identify the hidden variables and include them in the Pearson correction.

It is difficult to imagine any statistic that attempts to correct for hidden variables that does not make substantial assumptions about the form of the selection criteria and the distribution of the error terms. For these reasons, it seems that the best strategy is to simply use the Pearson correction formula. This matter is further addressed in section VII.

IV. DISCUSSION OF CONFIDENCE INTERVALS

The following theorem is well known. See, for example, Brunk (1960).

Theorem: If r is the sample correlation coefficient of a sample of size n from a bivariate normal population, then the statistic

$$z = \frac{1}{2} \log \frac{1 + r}{1 - r}$$

is asymptotically normally distributed with

$$E(z) = \frac{1}{2} \log \frac{1 + \rho}{1 - \rho}$$

and

$$V(z) \doteq \frac{1}{n-3},$$

where ρ is the population correlation coefficient being estimated by r and $\doteq$ means "approximately equal to." The approximation appears adequate for most purposes for sample sizes as small as 10.

If the distribution of the z-transform of the Pearson correction of r is also normal, then maybe, it would be the basis for the construction of interval estimates of ρ in the applicant population. Normality of the corrected statistic was tested with a chi-square test applied to data generated by CORR. Six cells were used for the χ^2 test. For the first test, the sample mean $(\overline{X})$ and the sample standard deviation (S) were used. In this case, the six cells were defined as $(-\infty, \overline{X} - 2S)$, $(\overline{X} - 2S, \overline{X} - S)$, $(\overline{X} - S, \overline{X})$, $(\overline{X}, \overline{X} + S)$, $(\overline{X} + 2S, \overline{X} + 2S)$, and $(\overline{X} + 2S, \infty)$. If the z-transform of the Pearson correction of r were normal, then the χ^2 values come from a chi-square distribution with 3 degrees of freedom. For a test at the $\alpha = .01$ level of significance, the critical value of χ^2 is 6.25. Based on eight runs with 3, 4, and 5 variables and various assumed correlation parameters, it seems likely that the corrected sample correlation coefficient has a distribution that is at least not significantly different from normal. The largest χ^2 observed in the eight runs was 6.17. On the eight runs, the χ^2 values were 1.44, 1.67, 6.17, 2.07, 0.74, 3.26, and 1.61. One more run was made with exactly the same parameters as the run that resulted in $\chi^2 = 6.17$, and for this run $\chi^2 = 2.42$.

There may, of course, be correlation parameters for which the distribution of the z-transform of the corrected statistic is not normal. Because of this problem, and because the standard deviation of the z-transform of the corrected statistic is greater than predicted by the theorem, in the next section, a procedure based on a simulation program is suggested. But first, it is necessary to determine if the mean of the z-transform of the corrected statistic is as predicted by the theorem, namely

$$(1) \qquad \frac{1}{2} \log \frac{1 + \rho}{1 - \rho}$$

Eight more runs were made with the same parameters as the previous eight. This time,

the hypothesis tested was that the distribution of the z-transform of the corrected statistic was normal with mean given by equation 1, the transform of the population parameter being estimated. Now the number of degrees of freedom of the χ^2 distribution is 4. For a test at the $\alpha = 0.1$ level of significance, the critical value is 7.78. On the eight runs, the observed χ^2 values were 1.90, 3.58, 1.93, 7.21, 2.41, 2.76, 5.41, and 2.02. For the example that gave $\chi^2 = 7.21$ another simulation was done and this time $\chi^2 = 0.71$.

Based on these observations, and others not presented here, it seems reasonable that in at least two aspects the z-transform of the corrected statistic behaves as the z-transform of the sample correlation coefficient. Namely, it is approximately normal, and the mean is the z-transform of the population parameter being estimated. The parameter being estimated in the case of the sample correlation coefficient is ρ^*, the correlation in the restricted population, and for the corrected statistic it is ρ, the correlation in the applicant population. The difficulty is that the standard deviation of the z-transform of the corrected statistic is larger than the standard deviation of the z-transform of the sample correlation coefficient. This last value is approximately $1/\sqrt{n-3}$, as predicted by the theorem. For the eight cases mentioned above, the standard deviation of the z-transform of the corrected statistic ranged from $1.05/\sqrt{n-3}$ *to* $1.7/\sqrt{n-3}$. So an interval estimate based on a standard deviation of $1/\sqrt{n-3}$ would be too small.

The standard deviation may depend on all of the input parameters. In the instance of the Air Force testing battery, there were 10 explicit selection variables and hence 10 variances and 45 correlation coefficients. Obviously, some way is needed to estimate the standard deviation of the z-transform of the corrected statistic to obtain an interval estimate of ρ. The only approach that seems feasible at this time is as follows.

Take the corrected correlations, the point estimates of the correlations in the applicant population, and do a simulation if these were the true population parameters. For the sample sizes typically used, there is every reason to believe that the standard deviation of the z-transform of the corrected statistic does not vary greatly with small changes in the population parameters. This is repeatedly observed in the simulation studies. The purpose of the simulation is to estimate the standard deviation of the z-transform of the corrected statistic, to calculate the chi-square value to test the normality assumption, and to print the 90%, 95%,

and the 99% capture ratios.

Let d represent the estimate of the standard deviation and

$$t = \frac{1}{2} \log \frac{1 + \rho}{1 - \rho},$$

where ρ is the correlation of interest in the applicant population. Then for each sample correlation coefficient (r) observed in the simulation, it is recorded whether or not it is true that

$$-z_{.05} \leq \frac{\frac{1}{2} \log \frac{1 + r}{1 - r}}{d} z_{.05}.$$

The fraction of time that this is true is called the 90% capture ratio. The 95% capture ratio is also computed, using $z_{.05}$. If the capture ratios are close to .90 and .95, respectively, and the chi-square value is small, then it is reasonable to compute a 90% or 95% confidence interval based on d. A derivation of this computation follows.

Let $(1 - \alpha) * 100\%$ be the confidence coefficient of the interval being defined. It is assumed that

$$1 - a = P(-z_{a/2} \leq \frac{\frac{1}{2} \log \frac{1 + r}{1 - r} - t}{d} \leq z_{a/2}).$$

After some algebraic manipulation, it is seen that this probability is the same as

$$(2) \qquad P\left(\frac{a - 1}{a + 1} \leq \rho \leq \frac{b - 1}{b + 1}\right)$$

where

$$a = \frac{1 r}{1 - r} e^{-2 z_{a/2} d}$$

and

$$b = \frac{1 + r}{1 - r} e^{2z_{\alpha/2}} d.$$

Thus, Equation 2 defines the (1 - a) * 100% confidence interval.

As an example of the use of this procedure, suppose that there are nine explicit selection variables called V_1, V_2,..., V_9 and one implicit selection variable V_{10}. $\rho_{V_1,V_{10}}$ is to be estimated using a sample of size 100. Use the correction formula to get point estimates of all 45 correlation coefficients as well as 10 variances. Suppose that the point estimate of $\rho_{V_1,V_{10}}$ is .5. Now plug all of these point estimates into the simulation program using the appropriate selection criteria. Suppose that the standard deviation of the z-transform of the correction (d) is 0.150. Note that $1/\sqrt{n-3} = .102$, so d is about 50% more than the standard deviation of the z-transform of $r_{V_1,V_{10}}$. Suppose that the capture ratio of the 90% and 95% confidence intervals are close to .90 and .95 respectively, and that the chi-square value is small. Hence

$$\frac{a - 1}{a + 1} = \frac{1.83 - 1}{1.83 + 1} = 0.29$$

where

$$a = \frac{1 + .5}{1 = .5} e^{-2(1.645)(.15)} = 1.83$$

and

$$\frac{b - 1}{b + 1} = \frac{4.914 - 1}{4.914 + 1} = 0.66$$

where

So the 90% confidence interval for ρ is (.29, .66). $b = \frac{1+1.5}{1-.5} e2(1.645)(.15) = 4.914$

V. TWO PROGRAMS FOR INTERVAL ESTIMATES

Two computer programs, NORM and TEST, have been written and implemented. Norm uses a multinormal distribution for the joint distribution of the test scores. The method of generating these multinormal observations is essentially the same as the method in program CORR and this method is presented in Jackson and Ree (1989). The only difference is that the selection process has been speeded up somewhat. TEST generates test scores by random sampling from a database of Air Force enlistee test scores. The main purpose of NORM is to estimate the standard deviation of the *z*-transform of the corrected sample correlation coefficient and to produce 90% and 95% confidence intervals based on the computations of the last section. The main purpose of TEST is to test this procedure using real test data.

The input and output for NORM is as described in Appendix A and B except that on output, more information is given on the *z*-transform. Both the 90% and 95% capture ratios are given as well as the chi-square value and the 90% and 95% confidence intervals. The meaning of these values and the intended use of program NORM is as follows.

Table 1 shows the corrected correlation coefficients based on some sample of 11 test scores, and an interval estimate is desired for the correlation between variables 1 and 11. The point estimate for this correlation is 0.596. The values in Table 1 are input to NORM as described in Appendix A, and a simulation is done as if these were the actual population parameters. Because these values are the best available estimates. The variable REPS represents the number of times an estimate of the correlation between variables 1 and 11 is calculated by the simulation. That is to say, it is the number of random samples that are generated, and for each random sample, the sample correlation coefficient and the corrected sample correlation coefficient are calculated from the data of that sample. Suppose we use REPS = 100. For each of the 100 repetitions, a 90%(95%) interval is constructed as explained in the previous section. The 90%(95%) capture ratio is the fraction of these 100 90%(95%) confidence intervals that contain the true value of 0.596. The chi-square value is

computed by the procedure presented in the last section to test the hypothesis that the z-transform of the corrected statistic is normal with mean

$$\frac{1}{2}\log\frac{1+.596}{1-.596} = .687.$$

Since this computation uses six cells and uses the sample standard deviation as the hypothesized standard deviation of the distribution, under the null hypothesis the computed value comes from a chi-square distribution with 4 degrees of freedom. For 4 degrees of freedom, the critical value for $\alpha = .1$ is 7.78 and for $\alpha = .5$ it is 9.49. Finally, the program computes a 90% and a 95% confidence interval using the equations of Section IV, the sample standard deviation for the 100 repetitions, and the value 0.596 as the point estimate. As stated in Section IV, these confidence intervals are based on the assumption that the transform of the corrected statistic is normal and unbiased. The purpose of the simulation is to estimate the standard deviation of the z-transform. The assumption is that the standard deviation obtained using the corrected estimates does not differ greatly from the true value of the standard deviation.

Table 1. Corrected data

1.000										
0.722	1.000									
0.801	0.708	1.000								
0.689	0.672	0.803	1.000							
0.524	0.627	0.617	0.608	1.000						
0.452	0.515	0.550	0.561	0.701	1.000					
0.637	0.533	0.529	0.423	0.306	0.225	1.000				
0.695	0.827	0.670	0.637	0.617	0.520	0.415	1.000			
0.695	0.684	0.593	0.521	0.408	0.336	0.741	0.600	1.000		
0.760	0.658	0.684	0.573	0.421	0.342	0.745	0.585	0.743	1.000	
0.596	0.749	0.487	0.489	0.465	0.433	0.503	0.680	0.640	0.570	1.000

After NORM has been started, it is just like running the program discussed in Appendix B. To start NORM type "run norm" at the VMS $ prompt. The results of the replicated calculations of the corrected (uncorrected) correlation coefficients are placed in NPLTC.DAT(NPLTU.DAT). Then type "run nplot" and respond as directed in Appendix B for program PLOT. One departure from the procedure outlined in Appendix B is that NORM terminates after the first plot. This is because the plot information is always placed in a file named PLT.PRT and so a second plot in the same run would cause the first to be overwritten before it could be examined. For several plots do "run nplot" several times. The program of Appendices A and B runs on a PC and is written in a version of PASCAL in which the pseudo random number generator must be supplied a seed to begin execution. This seed needs to be an odd integer and conventional wisdom holds that it should have about five digits for the best statistic properties of the stream of random numbers. The user is queried for the seed each time NORM executes. It is a good idea to run each simulation two or three times using a different seed each time.

The mean and variance of all variables in the model must be supplied to NORM in addition to the correlation coefficients of Table 1. For a given fixed subpopulation, the means and variances of the variables do not influence the outcome. The corrected values depend only on the correlations between the variables. Hence, any simulation can be done with variables all having mean zero and standard deviation one. So the following example includes only standard normal random variables. The program requires entry of the means and variances because if it just assumed that all variables were standard normal, then the user could not use the real selection criteria. One would, instead, have to derive a set of criteria that would produce the same subpopulation that the real criteria would have produced, had the actual means and variances of the variables been used.

A simulation was done using the coefficients in Table 1. These are the corrected values from a sample of Air Force recruits. All 11 variables were assumed to have mean zero and variance one (i.e., standard scores). The selected sample size was set at 200 and 100 replications were done. Thus, for each of the 100 samples generated, observations were made until 200 applicants met the selection criteria. As the resulting selection ratio for this example was about .25, this means that on the average about 800 observations were made for

each of the 100 samples generated. In the program, all variables must be given a name and in this case, the names given were X_1, X_2, ..., X_{11}. The first three variables were designated as explicit selection variables and the selection conditions were $X_1 \geq 0.2$, X_2, ≥ 0.2, and $X \geq 0.2$. The variables of interest were designated as X_1 and X_{11} and so the actual value of the correlation of interest is .596. After NORM and NPLOT have executed, the file PLT.PRT contains the summarized results of the simulation (see Figure 1).

Figure 1. PLT.PRT first run contents except that the histogram has been removed.

> Between X_1 and X_2 RHO = 0.596; a complete description is in malc.dat; the mean sel ratio is 0.254; the mean est is 0.599; the std-dev is 0.06; b_0 = 0.322 b_1 = 0.460; corrected; sample size is 200; # reps = 100 Smlest-Lrgest 0.451-0.744
>
> [THE PLOT GOES HERE]
>
> mean of the transform = 0.696; SD = 0.089; $1/sqrt(n - 3)$ = 0.071; 90% capture ratio = 0 9; 95% capture ration = 0.95; χ^2 = 2.164; 90% (0.494, 0.682); 95% (0.472, 0.697); back transform of the mean of the transform = 0.602

This file tells us that the correlation between the first and the eleventh variable, which is 0.596, is being estimated and that all the input parameters can be found in a file named MALC.DAT. Selection ratio is given, and it is stated that the sample mean of the 100 corrected estimates was 0.599. The sample standard deviation of the 100 corrected estimates was 0.06. The sample regression coefficients of X_{11} on X_1 are given, and it is noted that the values in this file pertain to the corrected statistic. Values for the uncorrected statistic may be obtained by running NPLOT and specifying the uncorrected option. By doing this, one would learn that, in this case, the sample mean of the 100 uncorrected estimates was 0.345. After the plot, omitted here, information about the z-transform of the 100 corrected values is presented. It is seen that the sample standard deviation of the transformed values is 0.089 and that the 90% and the 95% capture ratios are exactly as they should be. The small chi-

square value (2.164) indicates a good fit of the transformed values to a normal distribution with mean 0.687 (the *z*-transform of 0.596). Based on a standard deviation of 0.089, the 90% and 95% confidence intervals are (0.494, 0.682) and (0.472, 0.697), respectively. These interval estimates are associated with the point estimate 0.596. Also shown is $1/sqrt(n - 3) = 1/sqrt(197) = .071$. This is the expression given in the theorem at the beginning of the previous section as the approximate value of the standard deviation of the *z*-transform of the uncorrected statistic. In all of the simulation studies, this expression is a very good predictor of the observed sample standard deviation of the *z*-transform of the uncorrected statistic. For example, in the present example, this observed estimate was 0.074. The standard deviation of the *z*-transform of the corrected statistic is always greater than the standard deviation of the *z*-transform of the uncorrected statistic. However we must use the corrected statistic because the uncorrected statistic is frequently, as in the current example, very biased. For comparison, Figure 2 gives the results of another simulation using exactly the same input parameters but a different seed for the pseudo random number generator.

Figure 2. PLT.PRT second run but with different seed for the pseudo random number generator.

> Between X_1 and X_2 RHO = 0.596; a complete description is in malc.dat; the mean sel ratio is 0.253; the mean est is 0.593; the std-dev is 0.05; $b_0 = 0.319$ $b_1 = 0.455$; corrected; sample size is 200; # reps = 100 Smlest-Lrgest 0.427-0.718
>
> [THE PLOT GOES HERE]
>
> mean of the transform = 0.687; SD = 0.085; $1/sqrt(n - 3) = 0.071$; 90% capture ratio = 0.91; 95% capture ration = 0.97; $\chi^2 = 2.634$; 90% (0.498, 0.679); 95% (0.478, 0.693); back transform of the mean of the transform = 0.596

Program TEST is like program NORM except that the observed test scores are selected randomly from a database of test scores. If NREC represents the number of records in the

database file, then a random observation is made by selecting one number from 1, 2, ..., *NREC* and then reading that record. The record number is selected at random in such a way that the numbers 1, 2, ..., *NREC* all have an equal chance of being selected. This is done using a pseudo random number generator, so the user will be required to supply a seed each time the program runs. Sampling is done with replacement. In other words, a record may be selected more than once in a sample. This is, of course, not the way sampling is really done, but we are only interested in cases where NREC is sufficiently larger than the sample size and that sampling with replacement will have no noticeable effect on the outcome.

The input to TEST differs slightly from that of NORM. The user does not need to supply the means, variances, and correlation coefficients of all the random variables representing test scores, as these are implicit in the database. However, the correlation between the two variables of interest is requested so that it can be printed in the output. Also, the variables are not given names, but are referred to by the position they occupy in the records of the database file. The user will be prompted for the name of the database file (scores file) and the number of records in that file. The number of test scores in each record of the database file is, of course, the number of variables in the simulation. So the user is not required to specify the number of variables. That number is represented by a constant in the programs called on by TEST. In order to use TEST with a database file having a different number of test scores in each record, one program (TESTGBL) must be modified in one place and all programs must be compiled and linked again. The number is currently set to 11.

After TEST has been executed, TPLOT must be executed just as NPLOT must run after NORM. In order to calculate the capture ratio for the 90% and the 95% confidence intervals TPLOT uses the standard deviation from a corresponding run of NORM. A corresponding run means a run using the means, variances, and correlation coefficients of the variables in the database used by TEST. In this way, the capture ratios reflect the procedure discussed in the previous section. The output of TPLOT is left in file PLT.PRT, which is the same file in which NPLOT leaves its output. The output of TPLOT is almost identical to that of NPLOT. At the top of the output, the numbers of variables of interest is given along with the actual value of the correlation between these two variables.

VI. RUNS USING AIR FORCE TEST SCORE DATA

A set of test scores from Brooks AFB consists of 3,930 records, each containing 11 test scores. The means, standard deviations, and correlation coefficients of these data appear in Table 2.

These data have already been selected on the basis of the first 10 scores, the ASVAB. Variable 11 is the criterion variable. So Variables 1 through 10 are the explicit selection variables and variable 11 is the implicit selection variables. Notice that if variable 11 is linear and homoscedastic with respect to the first 10 variables before selection on those 10 variables has occurred, then it will be linear and homoscedastic after selection has occurred. Therefore, there is no logical difficulty with taking these data to represent an applicant population subject to more selection on the first 10 variables.

Table 2. Correlations of Tests and Criterion

Means										
82.54	53.12	53.78	54.08	55.19	54.94	54.86	52.34	52.89	54.46	52.16
Standard Deviations										
5.61	6.42	5.63	4.65	4.68	6.28	6.74	8.05	6.89	7.23	7.53
Correlation Coefficients										
1.00										
0.37	1.00									
0.31	0.23	1.00								
0.37	0.59	0.14	1.00							
0.34	0.37	0.17	0.43	1.00						
0.09	-0.08	0.25	-0.07	0.06	1.00					
0.08	-0.10	0.13	-0.03	0.10	0.54	1.00				
0.27	0.37	0.20	0.25	0.19	-0.16	-0.15	1.00			
0.28	0.31	0.57	0.20	0.20	0.27	0.14	0.05	1.00		
0.29	0.41	0.38	0.30	0.20	-0.06	-0.05	0.50	0.31	1.00	
0.32	0.49	0.24	0.38	0.24	-0.12	-0.12	0.56	0.20	0.49	1.00

A number of runs of TEST were made to try and answer two questions. The first is "How well does the confidence interval procedure described in Section IV work in this data

set?" Unfortunately, the answer is that it does not work well as would be liked. It produced a very reasonable capture ratio for the 90% and 95% confidence intervals in the majority of the cases ran, but sometimes the capture ratios were very low. In the worst case, presented below, the capture ratio for 90% was .73 and the capture ratio for 95% was .82. The conclusion, at least for this data set, was if one has no other way to get a feeling for the variability inherent in a sampling process, then this procedure is worthwhile. The second question is: "In this data set, was the corrected statistic more accurate than the uncorrected statistic?" The answer to this question is yes. The corrected statistic was too high and the uncorrected statistic was too low in every case ran. However, the distance between the true parameter value and the uncorrected estimate was typically two or three times the distance between the true parameter value and the corrected statistic. The conclusion is that one should always use the corrected statistic.

The two cases presented here are selected because they represent the most extreme cases encountered with respect to the second question above. The first case, where the corrected statistic was considerably better, is certainly more typical than the second case, where the corrected and uncorrected statistics are essentially the same distance from the true parameter value. The two cases also give the extreme values of the capture ratios for the 90% and 95% confidence intervals. With respect to the first question, neither case was typical. For the cases considered, the capture ratios bounced around between the two extreme values given in these two examples. In both cases, the sample size was 100 and the number of repetitions was 100.

In the first, there were three restrictions using all 10 explicit selection variables. The sum of the first three variables had to be at least 189.44, the sum of the next four had to be at least 219.07, and the sum of the last three explicit selection variables had to be at least 159.65. These three restrictions resulted in a selection ratio of 0.225. The correlation between variables 8 and 11, which can be seen from Table 14 to be .56, was the parameter estimated. The corrected estimate was .600 and the uncorrected estimate was .440. The capture ratios of the corrected statistic for the 90% and the 95% confidence intervals were 0.94 and 0.97, respectively.

In the second case, there was only one restriction. That restriction was that the sum of all 10 explicit selection variables had to be at least 568.2. With this one restriction, the selection ratio was 0.5. The correlation between variables 10 and 11, which was 0.49, was estimated. The corrected estimate was 0.558 and the uncorrected estimate was 0.425. The capture ratios of the corrected statistic for the 90% and the 95% confidence intervals were .73 and .82, respectively.

VII. RECOMMENDATIONS

Under the assumptions of linearity and homoscedasticity, Pearson correction does not change the value of the F statistic. If these assumptions are not satisfied, then the application of the correction formula is not appropriate. It should be noted that if the assumptions are not satisfied, linear regression is not appropriate. The application of the F test also requires normality of the error term for fixed X values and independence of error terms for different fixed X values. These assumptions probably do not hold exactly in the applicant population. But if they do hold in the applicant population, then they should also hold in the selected population provided the new test, whose predictive value is being tested, is not a hidden explicit selection variable. Range restriction may, of course, have an effect when real data are used, and there might be a fairly persistent bias in the F test one way or the other. If a large data set from the application population could be found, it would be advantageous to study the sampling distribution of the F statistic.

Section III establishes that hidden variables are potentially a severe problem. Assuming that linearity and homoscedasticity hold when all explicit selection variables are included, then they may not hold when an explicit selection is excluded from the model. This was demonstrated in Section III. If it can somehow be determined that hidden variables are sometimes causing inaccuracies in the Pearson correction statistic, then one must either include these variables in the correction or find some other statistic that gives correct results whether hidden variables are present or not. The application of such a statistic seems likely to require some assumption regarding the distribution of the error term in the linear model. Furthermore, this statistic is likely to not be very robust with respect to this assumption. For example, the statistic proposed by Gross and McGanney (1987) requires the assumption of normality of the error terms. It also requires that the selection criteria fit in one equation.

Finally, it has the property that it works poorly exactly when the Pearson correction statistic works well. These comments are not meant to be a criticism of the work of Gross and McGanney, but rather to point out the difficulty of developing a robust statistic that is able to detect the presence of hidden variables and correct for them. We should continue to seek a way to determine if hidden variables are a real problem. However, Pearson correction appears to work very well and should be replaced with another statistic only in the presence of overwhelming evidence of superior performance. The effect of range restriction is almost always present and Pearson correction does a very good job of correcting for this effect. If a large database of applicant scores, along with the values of suspected hidden variables for each record, could be obtained, then it would be possible to do an empirical study to determine if the inclusion of these variables would improve the accuracy of the Pearson correction statistic.

A procedure for interval estimates was described in Section IV, the programs to implement this procedure are described in Section V, and some empirical observations of the procedure were discussed in Section VI. It should be added to the observations of section VI that all of the runs made indicated that the sampling distribution of the corrected statistic is approximately normal even when the samples come from the database of test scores, as they do when running program TEST. It was also observed that the standard deviation of the z-transform of the corrected statistic for a corresponding run of NORM was a fairly good estimate of this standard deviation from a run of TEST. Thus, the reason that the capture ratios of the confidence intervals are sometimes too low is mostly because the corrected statistic from the database is frequently a slight overestimate of the true population parameter. Even if the confidence intervals were not entirely accurate, when decisions about the validity of tests must be made, it would be useful to have this information about the sampling distribution of corrected statistic. The capture ratios of confidence intervals based on the uncorrected statistic were very low because uncorrected statistic is considerably more of an underestimate than the corrected statistic was an overestimate.

References

Birnbaum, S.W., Paulson, E., & Andrews, F.C. (1950). On the effects of selection performed on some coordinates of a multi-dimensional population. Psychometrika, 15(2), 191-204.

Chatterjee & Price (1977). Regression analysis by example. John Wiley and Sons, 57-58.

Dunn, O.J. & Clark, V.A (1974). Applied Statistics: Analysis of variance and regression. John Wiley and Sons.

Gross, A.L. & McGanney, M.L. (1987). The restriction of range problem and nonignorable selection processes. Journal of Applied Psychology, 72(4), 604-610.

Jackson, D.E. & Ree, M.J. (1989). A tool for studying the effect of range restriction on correlation coefficient estimation.

Lawley, D.N (1943). A note on Karl Pearson's selection formulae. Proc. Roy. Soc. Edin, Sec A (Math & Phys.), 62, Part I, 28-30.

Pearson, Karl (1903). On the influence of natural selection on the variability and correlation of organs. Phil. Trans. Roy. Soc., A,200, 1-66.

APPENDIX A

General Description of the PC Simulation Program

The program was written in PASCAL and is currently running on an IBM compatible micro-computer. The joint distribution of all of the random variables is assumed to be multinormal in the unrestricted population. The inputs to the program are listed here for reference and they will be explained later as we discuss the program.

Figure A-1. Inputs to PASCAL program listed here for reference.

The number of variables [nv] and their names [vname]
Unrestricted population mean and std-dev [mu, sig]
The correlation coefficients in the unrestricted population [rho]
The number of explicitly selected variables [nve] the first nve entered
The number of restrictions [nr]
The coefficients of the explicitly selected variables [ncoeff]
Cutoff value for each restriction [cutoff]
Size of the unrestricted population [nwp]
Size of the restricted population [nvp]
The number of times the experiment will be repeated [reps]
The two variables of interest in the list of variables [int1, int2]
The number of hidden explicit selection variables [hes]
Cutoff for the hidden explicit selection variable [hcutoff]

Figure A-2 below is an example of a file describing the input to a run. The first line says that there are three variables in this case. The next three lines give the name, mean, and standard deviation of the three variables. In this case, they each have mean 0.0 and standard deviation 1.0. The next three lines give the correlation matrix for the three variables. So the correlation coefficient for (x,y) is 0.86, for (x,z) it is 0.0, and for (y,z) it is 0.43. The next line gives the number of explicit selection variables. There is one in this case and so X is the only explicit selection variable. Then it is specified that there is only one restriction (selection) and the restriction is $(1.0)X \geq 0.0$.

The selected group will consist of those persons getting a score of zero or greater on the X-test. Third to last line says that the variables of interest are 2 and 3 [Y and Z]. The second to last line says that there are no hidden variables. The other possibility is that this line could indicate that there was one hidden variable. If that were the case, then that variable is assumed to be the last variable, Z in this case. If one hidden variable were

specified, then the second number on this line would be the cutoff value for this variable. So if the second to last line were "1 0.00", then the selected group would consist of those

Figure A-2. An Input File

```
3
x      0.0    1.0
y      0.0    1.0
z      0.0    1.0
       1.00   0.86   0.0
       0.86   1.00   0.43
       0.00   0.43   1.0
1      # of explicitly restricted variables
1      number of restrictions
1.0    0.0
2      3 variables of interest
0      0.00 # hid exp sel vars cutoff value
1      50     100
```

persons getting a score of zero or greater on the X test and on the Z test. Data and a histogram of the distribution will be given for the uncorrected r between X and Z and the same information is given for the Pearson correction statistic. The program calculates the Pearson correction statistic using the theorem from Section I. The last line will be explained after the following discussion.

Creating a multinormal observation is equivalent to simulating one individual. In the above case this means getting three values, one for each of the three test scores X, Y, and Z. Each multinormal observation is part of the applicant group and is also a member of the selected group if the scores satisfy all of the restrictions. For the present case, this means that the score on the X test must be zero. One experiment is simulated by generating observations until two conditions are satisfied. There must be at least nwp observations in the applicant group and there must be at least nvp observations in the selected group. For must cases, we set nwp = 1 and then the only restriction is that we have at least nvp observations in the selected group. One run of the program consists of simulating reps experiments. The last line of a file which describes a run gives nwp, nvp, and reps in that order. In Figure A-2, nwp = 1, nvp = 50, and reps = 100.

When program CORR begins it will ask if the user wants to enter the data necessary to describe a run or to give the name of a file which contains the data in the expected format. The file in Figure A-2 is called test4 and so we can just give that name to CORR and the run is specified by the input parameters in Figure A-2. The reason test4 is in the expected format is because CORR wrote the file on a previous run. It was written when CORR executed, and it was specified that data would be entered from the keyboard and that these data were to be saved in a file named test4. Now if one is familiar with PASCAL read statements, they could use a text editor to change some of the parameters and use test4 for another run. After CORR executes, the data necessary to produce the histograms of the corrected and the uncorrected statistics are in two internal files and one must run program PLOT which will read these internal files and display this data on the printer.

For each experiment, CORR calculates each of the following quantities.

b0 and b1 = the estimates of the regression parameters.

statu = the uncorrected estimate of correlation coefficient.

statc = the corrected estimate of the correlation coefficient calculated with the equations of theorem 3.

Hence, CORR will generate reps copies of each of these parameters. In each case, the two implied variables are int1 and int2, and the regression parameters are for int2 on int1. In the case of b0 and b1, the only values retained are the totals of that after the reps experiments have been generated, the mean values of these parameters may be calculate. In the case of statu and statc, each observed value is retained and written to the files pltu.dat and pltc.dat, respectively. As mentioned earlier, the user can run PLOT to have all these results displayed.

Program CORR runs under the Turbo Pascal Version IV system on an IBM PC AT. It runs under the integrated environment Turbo.exe. This gives the user access to all of the features that make Turbo Pascal a user friendly system. For users not wanting to enter the integrated environment, there is a command line version explained in Chapter 12 of the Turbo Pascal Version IV owner's handbook. This Appendix, however, presupposes that the integrated environment is being used.

To begin the Turbo integrated environment type "turbo" followed by the "enter" key. Perhaps, it is preferable to keep the Turbo programs in one directory, say \v4, and the application programs in a subdirectory, say \v4\corr. Here, "application programs" means CORR plus all of the other programs and units that are part of the simulation system. In this subdirecory, one would also keep the data files that are input to the simulation programs. These data files describe the parameters of a run. For a complete description of the simulation programs and the format of the data files, see Appendix A. If the Turbo system is in \v4 and the applications are \v4\corr, then it might be useful to create in this subdirectory a file called "t.bat" that contained the single command "\v4\turbo." Then to start the system, get in directory \v4\corr and type "t" followed by the "enter" key.

Suppose that the Turbo programs and the application programs have all been loaded and the Turbo system has been initiated as discussed above. The screen is displaying the Turbo Pascal integrated environment main menu, consisting of File, Edit, Run, Compile, and Options. In normal circumstances, the only selections necessary to run simulations are File and Run. There are a number of ways to select menu options, but only one is mentioned here and it is the most general, in that it works from any place in the menu system. Simultaneously pressing the "alt" key and the first letter of any main menu option will select the option. To begin a simulation, select the File option by simultaneously pressing "alt" and "f." The screen now shows the File option. Press "1" to select the Load suboption. The screen responds with "*.pas." Hit the "enter" key and a directory of the application programs is shown. Use the arrow keys to place the highlighted rectangle over the name "corr.pas." Once the highlighted section is properly placed, hit the "enter" key. The screen shows the source listing of CORR and you are in the editor. Select the Run option by simultaneously pressing "alt" and "r." Now the simulation program begins to execute. While you are in the

editor, before selecting the Run option, it is important not to change the source code and if you do, then just do not save the changes and nothing will be damaged. From the editor, you can terminate Turbo Pascal by choosing the File option and then hitting "q" for quit.

Suppose that CORR has been started by loading it under the File option and then choosing the Run option as just explained. The first question asks if you want to describe the parameters for a new run (type "e" then the "enter" key) or if you want to specify the name of a data file that was created on a previous run and contains the parameter specifications (type "f" then the "enter" key). Since this is the first time we have executed the program, we select "e." Now the program is asking for the name of a file into which will be written the parameter specifications given in this run. Type a file name and then the "enter" key. Later, when the program is executed again, it will be possible to take the "f" option to the first question and then give this file name.

After having taken the "e" option to indicate that parameters will be entered from the keyboard, rather than from a file, you must give values for all of the input parameters discussed in Appendix A. Whenever giving a real number less than one, such as .543, Turbo Pascal insists that the number start with a zero. So you must enter 0.543 and not .543. In other words, Turbo Pascal wants one digit to precede the decimal point for all real numbers. Failure to comply with this rule will cause a runtime error and you will find yourself back in the editor. To begin again, select the File option, select the Load soboption, load CORR, then select the Run option and start over. After all parameters have been specified from the keyboard or from a specified file, the program will start doing the simulation. It prints a digit *d* when it has completed *d* tenths of the repetitions. In this way, it is possible to estimate, at any time, how much longer the run will take. There is no way to stop the program except at input/output operations. That is when the program is printing on the screen or waiting for input from the keyboard. If the "ctrl" key and the "pause" key are pressed simultaneously, then the program will be interrupted at the next input/output operation.

After the simulation is complete, you will be requested to strike any key to return to the Turbo Pascal menu. When you do strike a key, you will be back in the editor. Enter the File option and then the Load suboption as before and this time, place the highlighted area over the name "plot.pas" and press the "enter" key. Select the Run option and then program

PLOT will execute. At this point, the user must choose to see frequency data and sample statistics for the corrected statistic [c], the uncorrected statistic [u], or to terminate the program [q]. Suppose the "c" option is taken. Now the program is asking if the user wants to set the plot parameters or if certain default settings are desired. The usual choice here is "y" meaning "yes" the user wants the program to set the plot parameters. These parameters include the minimum value, the maximum value, the number of divisions, and the physical width of a division for the horizontal axis of the frequency distribution plot. The best strategy is to first let the program select these values and then it is possible to learn from the resulting plot the range of values taken by the simulated statistic. Then, one has some information to use in determining the optimal plot parameters and these may be specified on a subsequent pass. Next, the program wants to know if a plot of the statistic or the z-transform of the statistic is desired. If the z-transform is specified, then extra information about the sample statistics of the z-transformed observations is given after the plot.

The data at the top of the output from program PLOT is the same whether the z-transform is specified or not. It gives the names of the two variables of interest. It gives the true value of the correlation between the variables of interest in the applicant population. It gives the name of the file that contains the input parameters for the run. It also gives the mean selection ration, as well as the mean and standard deviation of the value of the corrected or uncorrected statistic These are calculated from the sample of "reps" calculations of the corrected or uncorrected statistic, each based on a sample of size "nvp." "Reps" and "nvp" are user supplied input parameters that describe the characteristics of a simulation. The selection ratio for one calculation is the fraction of observations generated that satisfied the selection criteria. The regression parameters (b_0 and b_1) for the second variable on the first are printed. It is indicated whether this information refers to the "corrected" or the "uncorrected" statistic. The sample size ("nvp") and the number of repetitions ("reps") are given. Finally, the smallest and largest values of the statistic or corrected statistic that were observed in the "reps" experiments are printed.

Next comes the histogram of the distribution of the uncorrected or the corrected statistic or of the z-transform of one of these. What is plotted here depends on what was

specified when program PLOT started to execute. On a later pass through this program, the scale of the horizontal axis may be changed.

If a *z*-transform plot was made, then extra information is printed after the histogram. The mean and standard deviation of the transform of the corrected or uncorrected statistic based on the "reps" repetitions is printed. The 95% capture ratio is given. This ratio is based on a confidence interval calculated using the standard deviation of the *z*-transform, not on $1/\sqrt{n-3}$. Finally, the inverse transform of the mean of the "reps" *z*-transform values is printed.